

# Reconstructing unrooted phylogenetic trees from symbolic ternary metrics

Stefan Grünewald

CAS-MPG Partner Institute for Computational Biology  
Chinese Academy of Sciences Key Laboratory of Computational Biology  
320 Yue Yang Road, Shanghai 200032, China  
Email: stefan@picb.ac.cn

Yangjing Long

Department of Mathematics  
Shanghai Jiao Tong University  
Dongchuan Road 800, Shanghai 200240, China  
Email: yjlong@sjtu.edu.cn

Yaokun Wu

Department of Mathematics and MOE-LSC  
Shanghai Jiao Tong University  
Dongchuan Road 800, Shanghai 200240, China  
Email: ykwu@sjtu.edu.cn

## Abstract

In 1998, Böcker and Dress gave a 1-to-1 correspondence between symbolically dated rooted trees and symbolic ultrametrics. We consider the corresponding problem for unrooted trees. More precisely, given a tree  $T$  with leaf set  $X$  and a proper vertex colouring of its interior vertices, we can map every triple of three different leaves to the colour of its median vertex. We characterise all ternary maps that can be obtained in this way in terms of 4- and 5-point conditions, and we show that the corresponding tree and its colouring can be reconstructed from a ternary map that satisfies those conditions. Further, we give an additional condition that characterises whether the tree is binary, and we describe an algorithm that reconstructs general trees in a bottom-up fashion.

**Keywords:** symbolic ternary metric ; median vertex ; unrooted phylogenetic tree

## 1 Introduction

A phylogenetic tree is a rooted or unrooted tree where the leaves are labeled by some objects of interest, usually taxonomic units (taxa) like species. The edges have a positive edge length, thus the tree defines a metric on the taxa set. It is a classical result in phylogenetics that the tree can be reconstructed from this metric, if it is unrooted or ultrametric. The latter means that the tree is rooted and all taxa are equally far away from the root. An ultrametric tree is realistic whenever the edge lengths represent the time and the taxa are species that can be observed in the present. In an ultrametric tree, the distance between two taxa is twice of the distance between each of the taxa and their last common ancestor (lca), hence pairs of taxa with the same lca must have the same distance. For three taxa  $x, y, z$ , it follows that there is no unique maximum within their three pairwise distances, thus we have  $d(x, y) \leq \max\{d(x, z), d(y, z)\}$ . This 3-point condition turns out to be sufficient for a metric to be ultrametric, too, and it is the key for reconstructing ultrametric trees from their distances. In 1995, Bandelt and Steel [1] observed that the complete ordering of the real numbers is not necessary to reconstruct trees, and they showed that the real-valued distances can be replaced by maps from the pairs of taxa into a cancellative abelian monoid. Later, Böcker and Dress [2] pushed this idea to the limit by proving that the image set of the symmetric map does not need any structure at all (see Section 2 for details). While this result is useful for understanding how little information it takes to reconstruct an ultrametric phylogenetic tree, it was not until recently that it turned out to have some practical applications. In 2013, Hellmuth et al. [5] found an alternative characterisation of symbolic ultrametrics in terms of cographs and showed that, for perfect data, phylogenetic trees can be reconstructed from orthology information. By adding some optimisation tools, this concept was then applied to analyse real data [6].

Motivated by the practical applicability of symbolic ultrametrics, we are considering their unrooted version. However, in an unrooted tree there is in general no interior vertex associated to a pair of taxa

that would correspond to the last common ancestor in a rooted tree. Instead, there is a median associated to every set of three taxa that represents, for every possible rooting of the tree, a last common ancestor of at least two of the three taxa. Therefore, we consider ternary maps from the triples of taxa into an image set without any structure. We will show that an unrooted phylogenetic tree with a proper vertex colouring can be reconstructed from the function that maps every triple of taxa to the colour of its median.

In order to apply our results to real data, we need some way to assign a state to every set of three taxa, with the property that 3-sets with the same median will usually have the same state. For symbolic ultrametrics, the first real application was found 15 years after the development of the theory. In addition to the hope that something similar happens with symbolic ternary metrics, we have some indication that they can be useful to construct unrooted trees from orthology relations (see Section 6 for details).

Consider an unrooted tree  $T$  with vertex set  $V$ , edge set  $E$ , and leaf set  $X$ , and a dating map  $t : V \rightarrow M^\odot$ , where  $M^\odot = M \cup \{\odot\}$  such that  $t(x) = \odot$  for all  $x \in X$ , and  $t(v_1) \neq t(v_2)$  if  $v_1 v_2 \in E$ .

For any  $S = \{x, y\} \in \binom{X}{2}$  there is a unique path  $[x, y]$  with end points  $x$  and  $y$ , and for any 3-set  $S = \{x, y, z\} \in \binom{V}{3}$  there is a unique *triple point* or *median*  $\text{med}(x, y, z)$  such that  $[x, y] \cap [y, z] \cap [x, z] = \{\text{med}(x, y, z)\}$ . Putting  $[x, x] = \{x\}$ , the definition also works, if some or all of  $x, y$  and  $z$  equal.

Given a phylogenetic tree  $T$  on  $X$  and a dating map  $t : V \rightarrow M^\odot$ , we can define the symmetric symbolic ternary map  $d_{(T;t)} : X \times X \times X \rightarrow M^\odot$  by  $d_{(T;t)}(x, y, z) = t(\text{med}(x, y, z))$ .

In this set-up, our question can be phrased as follows: Given an arbitrary symbolic ternary map  $\delta : X \times X \times X \rightarrow M^\odot$ , can we determine if there is a pair  $(T; t)$  for which  $d_{(T;t)}(x, y, z) = \delta(x, y, z)$  holds for all  $x, y, z \in X$ ?

The rest of this paper is organized as follows. In Section 1.1, we present the basic and relevant concepts used in this paper. In Section 2 we recall the one-to-one correspondence between symbolic ultrametrics and symbolically dated trees, and introduce our main results Theorem 4 and Theorem 5.

In Section 3 we give the proof of Theorem 4. In order to prove our main result, we first introduce the connection between phylogenetic trees and quartet systems on  $X$  in Subsection 3.1. Then we use a graph representation to analyse all cases of the map  $\delta$  for 5-taxa subsets of  $X$  in Subsection 3.2.

In Section 4 we use a similar method to prove Theorem 5, which gives a sufficient and necessary condition to reconstruct a binary phylogenetic tree on  $X$ .

In Section 5, we give a criterion to identify all pseudo-cherries of the underlying tree from a symbolic ternary metric. This result makes it possible to reconstruct the tree in a bottom-up fashion.

In the last section we discuss some open questions and future work.

## 1.1 Preliminaries

We introduce the relevant basic concepts and notation. Unless stated otherwise, we will follow the monographs [7] and [4].

In the remainder of this paper,  $X$  denotes a finite set of size at least three.

An (*unrooted*) *tree*  $T = (V, E)$  is an undirected connected acyclic graph with vertex set  $V$  and edge set  $E$ . A vertex of  $T$  is a *leaf* if it is of degree 1, and all vertices with degree at least two are *interior* vertices.

A *rooted tree*  $T = (V, E)$  is a tree that contains a distinguished vertex  $\rho_T \in V$  called the *root*. We define a partial order  $\preceq_T$  on  $V$  by setting  $v \preceq_T w$  for any two vertices  $v, w \in V$  for which  $v$  is a vertex on the path from  $\rho_T$  to  $w$ . In particular, if  $v \preceq_T w$  and  $v \neq w$  we call  $v$  an *ancestor* of  $w$ .

An unrooted *phylogenetic tree*  $T$  on  $X$  is an unrooted tree with leaf set  $X$  that does not contain any vertex of degree 2. It is *binary*, if every interior vertex has degree 3.

A rooted phylogenetic tree  $T$  on  $X$  is a rooted tree with leaf set  $X$  that does not contain any vertices with in- and out-degree one, and whose root  $\rho_T$  has in-degree zero. For a set  $A \subseteq X$  with cardinality at least 2, we define the *last common ancestor* of  $A$ , denoted by  $\text{lca}_T(A)$ , to be the unique vertex in  $T$  that is the greatest lower bound of  $A$  under the partial order  $\preceq_T$ . In case  $A = \{x, y\}$  we put  $\text{lca}_T(x, y) = \text{lca}_T(\{x, y\})$ .

Given a set  $Q$  of four taxa  $\{a, b, c, d\}$ , there exist always exactly three partitions into two pairs:  $\{\{a, b\}, \{c, d\}\}, \{\{a, c\}, \{b, d\}\}$  and  $\{\{a, d\}, \{b, c\}\}$ . These partitions are called *quartets*, and they represent the three non-isomorphic unrooted binary trees with leaf set  $Q$ . These trees are usually called quartet trees, and they – as well as the corresponding quartets – are symbolised by  $ab|cd, ac|bd, ad|bc$  respectively. We use  $Q(X)$  to denote the set of all quartets with four taxa in  $X$ . A phylogenetic tree  $T$  on  $X$  *displays* a quartet  $ab|cd \in Q(X)$ , if the path from  $a$  to  $b$  in  $T$  is vertex-disjoint with the path from  $c$  to  $d$ . The collection of all quartets that are displayed by  $T$  is denoted by  $Q_T$ .

Let  $M$  be a non-empty finite set,  $\odot$  denotes a special element not contained in  $M$ , and  $M^\odot := M \cup \{\odot\}$ . Note that in biology the symbol  $\odot$  corresponds to a “non-event” and is introduced for purely technical reasons [5]. A *symbolic ternary map* is a mapping from  $X \times X \times X$  to  $M^\odot$ . Suppose we have a symbolic ternary map  $\delta : X \times X \times X \rightarrow M^\odot$ , we say  $\delta$  is *symmetric* if the value of  $\delta(x, y, z)$  is only related to the set  $\{x, y, z\}$  but not on the ordering of  $x, y, z$ , i.e., if  $\delta(x, y, z) = \delta(y, x, z) = \delta(z, y, x) = \delta(x, z, y)$  for all  $x, y, z \in X$ . For simplicity, if a map  $\delta : X \times X \times X \rightarrow M^\odot$  is symmetric, then we can define  $\delta$  on the set  $\{x, y, z\}$  to be  $\delta(x, y, z)$ .

Let  $S$  be a set, we define  $|S|$  to be the number of elements in  $S$ .

## 2 Symbolic ultrametrics and our main results

In this section, we first recall the main result concerning symbolic ultrametrics by Böcker and Dress [2].

Suppose  $\delta : X \times X \rightarrow M^\odot$  is a map. We call  $\delta$  a *symbolic ultrametric* if it satisfies the following conditions:

- (U1)  $\delta(x, y) = \odot$  if and only if  $x = y$ ;
- (U2)  $\delta(x, y) = \delta(y, x)$  for all  $x, y \in X$ , i.e.,  $\delta$  is symmetric;
- (U3)  $|\{\delta(x, y), \delta(x, z), \delta(y, z)\}| \leq 2$  for all  $x, y, z \in X$ ; and
- (U4) there exists no subset  $\{x, y, u, v\} \in \binom{X}{4}$  such that  $\delta(x, y) = \delta(y, u) = \delta(u, v) \neq \delta(y, v) = \delta(x, v) = \delta(x, u)$ .

Now suppose that  $T = (V, E)$  is a rooted phylogenetic tree on  $X$  and that  $t : V \rightarrow M^\odot$  is a map such that  $t(x) = \odot$  for all  $x \in X$ . We call such a map  $t$  a *symbolic dating map* for  $T$ ; it is *discriminating* if  $t(u) \neq t(v)$ , for all edges  $\{u, v\} \in E$ . Given  $(T, t)$ , we associate the map  $d_{(T, t)}$  on  $X \times X$  by setting, for all  $x, y \in X$ ,  $d_{(T, t)}(x, y) = t(\text{lca}_T(x, y))$ . Clearly  $\delta = d_{(T, t)}$  satisfies Conditions (U1), (U2), (U3), (U4) and we say that  $(T, t)$  is a *symbolic representation* of  $\delta$ . Böcker and Dress established in 1998 the following fundamental result which gives a 1-to-1 correspondence between symbolic ultrametrics and symbolic representations [2], i.e., the map defined by  $(T, t) \mapsto d_{(T, t)}$  is a bijection from the set of symbolically dated trees into the set of symbolic ternary metrics.

**Theorem 1** (Böcker and Dress 1998 [2]). *Suppose  $\delta : X \times X \rightarrow M^\odot$  is a map. Then there is a discriminating symbolic representation of  $\delta$  if and only if  $\delta$  is a symbolic ultrametric. Furthermore, up to isomorphism, this representation is unique.*

Similarly, we consider unrooted trees. Suppose that  $T = (V, E)$  is an unrooted tree on  $X$  and that  $t : V \rightarrow M^\odot$  is a symbolic dating map, i.e.,  $t(x) = \odot$  for all  $x \in X$  and  $t(x) \neq t(y)$  if  $(x, y) \in E$ . Given the pair  $(T, t)$ , we associate the map  $\delta_{(T, t)}$  on  $X \times X \times X$  by setting, for all  $x, y, z \in X$ ,  $\delta_{(T, t)}(x, y, z) = t(\text{med}(x, y, z))$ .

Before stating our main results, we need the following definition:

**Definition 2** (*n-m partitioned*). *Suppose  $\delta : X \times X \times X \rightarrow M^\odot$  is a symmetric map. We say that a subset  $S$  of  $X$  is *n-m partitioned* (by  $\delta$ ), if among all the 3-element subsets of  $S$ , there are in total 2 different values of  $\delta$ , and  $n$  of those 3-sets are mapped to one value while all other  $m$  3-sets are mapped to the other value.*

Note that  $S$  can be *n-m partitioned*, only when  $\binom{|S|}{3} = m + n$ .

**Definition 3** (*symbolic ternary metrics*). *We say  $\delta : X \times X \times X \rightarrow M^\odot$  is a symbolic ternary metric, if the following conditions hold.*

- (1)  $\delta$  is symmetric, i.e.,  $\delta(x, y, z) = \delta(y, x, z) = \delta(z, y, x) = \delta(x, z, y)$  for all  $x, y, z \in X$ .
- (2)  $\delta(x, y, z) = \odot$  if and only if  $x = z$  or  $y = z$  or  $x = y$ .
- (3) for any distinct  $x, y, z, u$  we have

$$|\{\delta(x, y, z), \delta(x, y, u), \delta(x, z, u), \delta(y, z, u)\}| \leq 2,$$

and when the equality holds then  $\{x, y, z, u\}$  is 2-2 partitioned by  $\delta$ .

- (4) there is no distinct 5-element subset  $\{x, y, z, u, e\}$  of  $X$  which is 5-5 partitioned by  $\delta$ .

We will refer to these conditions throughout the paper.

Our main result is:

**Theorem 4.** *There is a 1-to-1 correspondence between the symbolically dated phylogenetic trees and the symbolic ternary metrics on  $X$ .*

Let  $\delta$  be a ternary symbolic ultrametric on  $X$ . Then we call  $\delta$  *fully resolved*, if the following condition holds:

- (\*) If  $|\{\delta(x, y, z), \delta(x, y, u), \delta(x, z, u), \delta(y, z, u)\}| = 1$ , then there exists  $e \in X$  such that  $e$  can resolve  $xyz u$ . i.e., the set  $\{x, y, z, u, e\}$  is 4-6 partitioned by  $\delta$ .

Now we can characterise ternary symbolic ultrametrics that correspond to binary phylogenetic trees:

**Theorem 5.** *There is a 1-to-1 correspondence between the symbolically dated binary phylogenetic trees and the fully resolved symbolic ternary metrics on  $X$ .*

## 3 Reconstructing a symbolically dated phylogenetic tree

The aim of this section is to prove Theorem 4.

### 3.1 Quartet systems

We will use quartet systems to prove Theorem 4. In 1981, Colonius and Schulze [3] found that, for a quartet system  $Q$  on a finite taxa set  $X$ , there is a phylogenetic tree  $T$  on  $X$  such that  $Q = Q_T$ , if and only if certain conditions on subsets of  $X$  with up to five elements hold. The following theorem (Theorem 3.7 in [4]) states their result.

A quartet system  $Q$  is *thin*, if for every 4-subset  $a, b, c, d \subseteq X$ , at most one of the three quartets  $ab|cd$ ,  $ac|bd$  and  $ad|bc$  can be contained in  $Q$ . It is *transitive*, if for any 5 distinct elements  $a, b, c, d, e \in X$ , the

quartet  $ab|cd$  is in  $Q$  whenever both of the quartets  $ab|ce$  and  $ab|de$  are contained in  $Q$ . It is *saturated*, if for any five distinct elements  $a, b, c, d, e \in X$  with  $ab|cd \in Q$ , at least one of the two quartets  $ae|cd$  and  $ab|ce$  is also in  $Q$ .

**Theorem 6.** *A quartet system  $Q \subseteq Q(X)$  is of the form  $Q = Q(T)$  for some phylogenetic tree  $T$  on  $X$  if and only if  $Q$  is thin, transitive and saturated.*

We can encode a phylogenetic tree on  $X$  in terms of a quartet system by taking all the quartets displayed by the tree, as two phylogenetic trees on  $X$  are isomorphic if and only if the associated quartet systems coincide [4]. Hence, a quartet system that satisfies Theorem 6 uniquely determines a phylogenetic tree.

### 3.2 Graph representations of a ternary map

Suppose we have a symmetric map  $\delta : X \times X \times X \rightarrow M^\odot$ . Then we can represent the restriction of  $\delta$  to all 3-element subsets of any 5-element subset  $\{x, y, z, u, v\}$  of  $X$  by an edge-coloured complete graph on the 5 vertices  $x, y, z, u, v$ . For any distinct  $a, b, c, d \in \{x, y, z, u, v\}$ , edge  $ab$  and edge  $cd$  have the same colour if and only if the value of  $\delta$  for  $\{x, y, z, u, v\} \setminus \{a, b\}$  is the same as for  $\{x, y, z, u, v\} \setminus \{c, d\}$ . It follows from Condition (3) in the definition of a symbolic ternary metric that, for any vertex of the graph, either 2 incident edges have one colour and the other 2 edges have another colour, or all 4 incident edges have the same colour. By symmetry, there are exactly five non-isomorphic graph representations.

**Lemma 1.** *Let the edges of a  $K_5$  be coloured such that for each vertex, the 4 incident edges are either coloured by the same colour, or 2 of them coloured by one colour and the other 2 by another colour. Then there are exactly 5 non-isomorphic colourings, and they are depicted in Figure 1.*

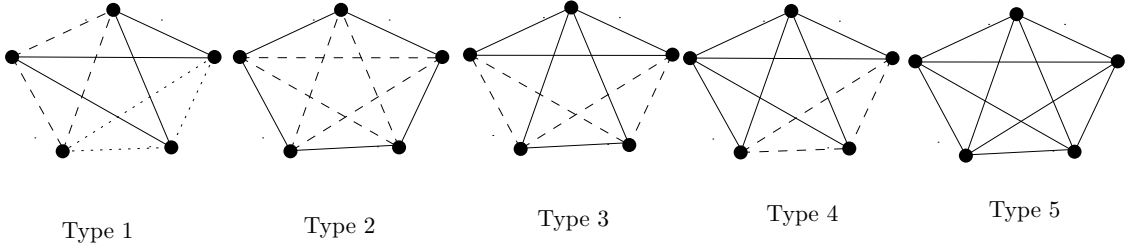


Figure 1: The 5 non-isomorphic colourings of  $K_5$  for which every colour class induces an Eulerian graph.

*Proof.* It follows from the condition on the colouring that every colour class induces an *Eulerian* subgraph (a graph where all vertices have even degree) of  $K_5$ . Therefore, ignoring isolated vertices, every such induced subgraph either is a cycle or it contains a vertex of degree four. Since there are only ten edges, the only way to have three colour classes is two triangles and one 4-cycle. In that case each of the triangles must contain two non-adjacent vertices of the 4-cycle and the vertex that is not in the 4-cycle, thus we get a colouring isomorphic to Type 1 in Figure 1. If there are exactly two colour classes, then one of them has to be a cycle and the other one its complement. This yields Types 2, 3, and 4, if the length of the cycle is 5, 4, and 3, respectively. Finally, if there is only one colour, we get Type 5.  $\square$

Note that the vertices are not labeled and it does not matter which colour we are using.

We will prove Theorem 4 by obtaining a quartet system from any symbolic ternary metric. More precisely, we say that the symbolic ternary metric  $\delta$  on  $X$  *generates* the quartet  $xy|zu$  if either  $\delta(x, z, u) = \delta(y, z, u) \neq \delta(x, y, z) = \delta(x, y, u)$ , or  $|\{\delta(x, y, z), \delta(x, y, u), \delta(x, z, u), \delta(y, z, u)\}| = 1$  and there is  $e \in X$  such that

$$\begin{aligned} \delta(x, y, e) &= \delta(x, y, z) = \delta(x, y, u) = \delta(x, z, u) = \delta(z, u, e) = \delta(y, z, u) \\ &\neq \delta(x, u, e) = \delta(x, z, e) = \delta(y, z, e) = \delta(y, u, e). \end{aligned}$$

In the latter case, we say that  $e$  *resolves*  $x, y, z, u$ . Note that the 3-sets obtained by adding  $e$  to the pairs of the generated quartet both have the same  $\delta$ -value as the subsets of  $\{x, y, z, u\}$ . The following lemma will show that the set of all quartets generated by a symbolic ternary metric is thin.

**Lemma 2.** *Let  $\delta : X \times X \times X \rightarrow M^\odot$  be a symbolic ternary metric and let  $x, y, z, u \in X$  be four different taxa with  $|\{\delta(x, y, z), \delta(x, y, u), \delta(x, z, u), \delta(y, z, u)\}| = 1$ . Let  $e, e' \in X - \{x, y, z, u\}$  such that  $\{x, y, z, u, e\}$  and  $\{x, y, z, u, e'\}$  are both 4-6-partitioned, and let*

$$\begin{aligned} \delta(x, y, e) &= \delta(x, y, z) = \delta(x, y, u) = \delta(x, z, u) = \delta(z, u, e) = \delta(y, z, u) \\ &\neq \delta(x, u, e) = \delta(x, z, e) = \delta(y, z, e) = \delta(y, u, e). \end{aligned}$$

Then we also have

$$\begin{aligned} \delta(x, y, e') &= \delta(x, y, z) = \delta(x, y, u) = \delta(x, z, u) = \delta(z, u, e') = \delta(y, z, u) \\ &\neq \delta(x, u, e') = \delta(x, z, e') = \delta(y, z, e') = \delta(y, u, e'). \end{aligned}$$

*Proof.* Suppose the statement of the lemma is wrong. Then we can assume without loss of generality that

$$\begin{aligned} \delta(x, z, e') &= \delta(x, y, z) = \delta(x, z, u) = \delta(y, u, e') = \delta(x, y, u) = \delta(y, z, u) \\ &\neq \delta(x, u, e') = \delta(x, y, e') = \delta(y, z, e') = \delta(z, u, e'). \end{aligned}$$

Case (a):  $\delta(x, u, e) \neq \delta(x, u, e')$ . We assume that  $\delta(x, u, e)$  is dashed,  $\delta(x, u, e')$  is dotted, and  $\delta(x, y, z)$  is solid. Since  $\delta$  is a symbolic ternary metric, by Lemma 1, the graph representation of  $\{y, z, u, e, e'\}$  has to be Type 1, so the colour classes are one 4-cycle and two 3-cycles. The values of  $\delta$  for the sets that contain at most one of  $e$  and  $e'$  are shown in Figure 2. There is a path of length 3 that is coloured with  $\delta(x, y, z)$  (solid), and there are paths of length 2 coloured with  $\delta(x, u, e)$  (dashed) and  $\delta(x, u, e')$  (dotted), respectively. It follows that we only can get Type 1 by colouring the edges connecting the end vertices of each of those paths with the same colour as the edges on the path. We get  $\delta(u, e, e') = \delta(x, y, z)$  (solid),  $\delta(z, e, e') = \delta(x, u, e')$  (dotted), and  $\delta(y, e, e') = \delta(x, u, e)$  (dashed). Now doing the same analysis for  $\{x, y, z, e, e'\}$  yields  $\delta(z, e, e') = \delta(x, u, e)$ , in contradiction to  $\delta(z, e, e') = \delta(x, u, e')$ .

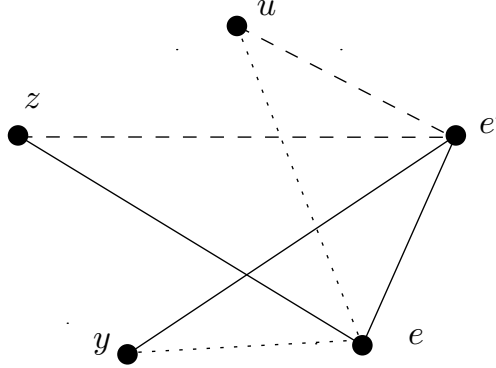


Figure 2: The partial colouring of  $K_5$  as described in Case (a).

Case (b):  $\delta(x, u, e) = \delta(x, u, e')$ . The graph representation of  $\{y, z, u, e, e'\}$  can be obtained from Figure 2 by identifying the colours dashed and dotted. It contains a path of length 3 of edges coloured with  $\delta(x, y, z)$  and a path of length 4 of edges coloured with  $\delta(x, u, e)$ . Since a path is not an Eulerian graph, both colours must be used for at least one of the remaining three edges, thus Type 1 is not possible. Hence,  $\{y, z, u, e, e'\}$  must be 4-6-partitioned, and we get  $\delta(u, e, e') = \delta(x, y, z)$ ,  $\delta(y, e, e') = \delta(z, e, e') = \delta(x, u, e)$ . Now we consider the graph representation of  $\{x, z, u, e, e'\}$ , and we observe that the edges coloured with  $\delta(x, y, z)$  contain a path of length 4, and the edges coloured with  $\delta(x, u, e)$  contain a 5-cycle. Hence,  $\{x, z, u, e, e'\}$  is 5-5-partitioned, in contradiction to Condition (4).  $\square$

*Proof of Theorem 4.* By the definitions of the median, the ternary map  $\delta_{(T;t)}$  associated with a symbolically dated phylogenetic tree  $(T, t)$  on  $X$  satisfies Conditions (1) and (2). For any distinct leaves  $x, y, z, u$ , the smallest subtree of  $T$  connecting those four leaves has at most two vertices of degree larger than two. Hence,  $\{\text{med}(x, y, z), \text{med}(x, y, u), \text{med}(x, z, u), \text{med}(y, z, u)\}$  consists of at most two distinct median points, so there are at most 2 distinct values of  $\delta_{(T;t)}$ . Therefore,  $\delta_{(T;t)}$  satisfies Condition (3).

For any 5 distinct leaves, the smallest subtree of  $T$  connecting them either has three vertices of degree 3, or one vertex of degree 3 and one of degree 4, or one vertex of degree 5, while all other vertices have degree 1 or 2. The first case is depicted in Figure 3. There  $v_1$  is the median for the 3-sets that contain  $x_1$  and  $x_2$ ,  $v_3$  is the median for the 3-sets that contain  $z_1$  and  $z_2$ , and  $v_2$  is the median for the remaining four 3-sets. Hence, either  $\{x_1, x_2, y, z_1, z_2\}$  is 4-6-partitioned (if  $t(v_1) = t(v_3)$ ), or there are three different values of  $\delta_{(T;t)}$  within those five taxa. For the other two cases, the set of five taxa is either 3-7-partitioned or  $\delta_{(T;t)}$  is constant on all its subsets with 3 taxa. Hence, no subset of  $X$  of cardinality five is 5-5-partitioned by  $\delta_{(T;t)}$ , thus  $\delta_{(T;t)}$  satisfies Condition (4).

On the other hand, let  $\delta$  be a symbolic ternary metric on  $X$ . By Lemma 1, taking any 5-element subset of  $X$ , the possible graph representations of the delta system satisfying (1), (2), and (3) are shown in Figure 1. Except for Type 2, all other types satisfy (4).

For the first type, the delta system is  $\delta(y, z, u) = \delta(x, y, z) = \delta(w, y, z) \neq \delta(w, x, z) = \delta(w, x, u) = \delta(w, x, y) \neq \delta(x, z, u) = \delta(w, z, u) = \delta(x, y, u) = \delta(w, y, u) \neq \delta(y, z, u)$ . The corresponding quartet system is  $\{xw|yu, xw|zu, xw|yz, uw|yz, xu|yz\}$ .

For the third type, the delta system is  $\delta(y, z, u) = \delta(x, z, u) = \delta(w, z, u) = \delta(x, y, w) = \delta(y, u, w) = \delta(y, z, w) \neq \delta(x, y, z) = \delta(w, x, z) = \delta(x, y, u) = \delta(w, x, u)$ . The corresponding quartet system is  $\{wy|xu, wy|xz, xy|zu, wy|zu, wx|zu\}$ .

For the fourth type, the delta system is  $\delta(w, x, z) = \delta(w, x, u) = \delta(w, x, y) \neq \delta(y, z, u) = \delta(x, y, z) = \delta(x, z, u) = \delta(w, z, u) = \delta(w, y, z) = \delta(x, y, u) = \delta(w, y, u)$ . The corresponding quartet system is  $\{xw|yu, xw|zu, xw|yz\}$ .

For the fifth type, the delta system is  $\delta(y, z, u) = \delta(x, y, z) = \delta(w, y, z) = \delta(w, x, z) = \delta(w, x, u) = \delta(w, x, y) = \delta(x, z, u) = \delta(w, z, u) = \delta(x, y, u) = \delta(w, y, u)$ . The corresponding quartet system is  $\emptyset$ .

All quartet systems are thin, transitive, and saturated. Indeed, the delta systems of Types 1 and 3 generate all quartets displayed by a binary tree, Type 4 generates all quartets displayed by a tree with exactly one interior edge, and Type 5 corresponds to the star tree with 5 leaves. Now we take the union of all quartets generated by  $\delta$ . The resulting quartet system is transitive and saturated, since we have verified these properties for all subsets of five taxa, and it is thin in view of Lemma 2. Therefore, by Theorem 6, every delta system satisfying Conditions (1), (2), (3) and (4) uniquely determines a phylogenetic tree  $T$  on  $X$ .

It only remains to show that two 3-element subsets of  $X$  that have the same median in  $T$  must be mapped to the same value of  $\delta$ , since then we can define  $t$  to be the dating map with  $t(v) = \delta(x, y, z)$  for every interior vertex  $v$  of  $T$  and for all 3-sets  $\{x, y, z\}$  whose median is  $v$ . It suffices to consider two sets which intersect in two taxa, as the general case follows by exchanging one taxon up to three times. We assume that  $v$  is the median of both,  $\{w, x, y\}$  and  $\{w, x, z\}$ . If  $\delta(w, x, y) \neq \delta(w, x, z)$ , then  $\delta$  generates one of the quartets  $wy|xz$  and  $wz|xy$ , thus  $T$  must display that quartet. However, in both cases  $\{w, x, y\}$  and  $\{w, x, z\}$  do not have the same median in  $T$ .  $\square$

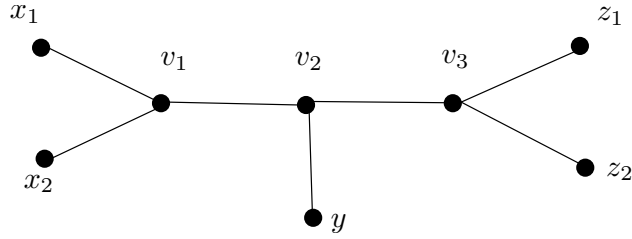


Figure 3: The leaves and median vertices for a 5-taxa binary tree.

The set of quartets generated by Type 2 does not satisfy the condition of being saturated from Theorem 6. Without loss of generality, label the vertices by  $x, y, z, u, w$  as in Figure 4. Then the quartet system is  $\{yw|zu, xu|yz, xz|uw, xy|zw, xw|yu\}$ . In order to be saturated, the presence of  $xu|yz$  would induce that we have  $xw|yz$  or  $xu|yw$ , but we have  $xy|zw, xw|yu$  instead.

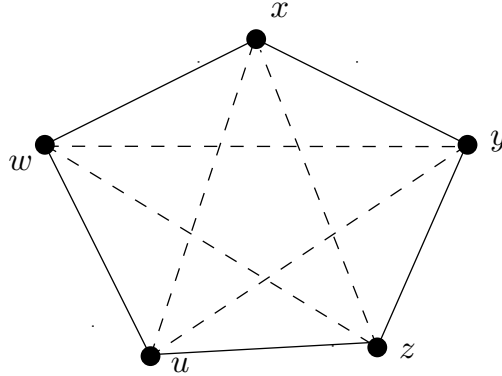


Figure 4: The graph representation of a ternary map satisfying Conditions (1), (2), and (3), but not (4).

## 4 Reconstructing a binary phylogenetic tree

The aim of this section is to prove Theorem 5.

A quartet system  $Q$  on  $X$  is *complete*, if

$$|\{Q \cap \{ab|cd, ac|bd, ad|bc\}\}| = 1$$

holds for all  $\{a, b, c, d\} \in \binom{X}{4}$ . Using the easy observation that a phylogenetic tree is binary if and only if it displays a quartet for every 4-set, the following result is a direct consequence of Theorem 6.

**Corollary 1.** *A quartet system  $Q \subseteq Q(X)$  is of the form  $Q = Q(T)$  for some binary phylogenetic tree  $T$  on  $X$  if and only if  $Q$  is complete, transitive, and saturated.*

Condition (\*) ensures that a ternary metric  $\delta$  generates a quartet for every set of four taxa, even if  $\delta$  is constant on all of its 3-taxa subsets. In view of Lemma 2 we also have that  $\delta$  can not generate two different quartets for the same 4-set. Hence, we have the following corollary.

**Corollary 2.** *A symbolic ternary metric  $\delta : X \times X \times X \rightarrow M^\odot$  that satisfies Condition (\*) generates a complete quartet system on  $X$ .*

Now we prove Theorem 5.

*Proof.* Let  $(T, t)$  be a symbolically dated binary phylogenetic tree. By Theorem 4,  $\delta_{(T; t)}$  is a symbolic ternary metric. Since  $T$  is binary, it displays a quartet for every 4-taxa subset  $\{x, y, z, u\}$  of  $X$ . Assume that  $T$  displays  $xu|yz$ , thus  $\text{med}(x, y, z) \neq \text{med}(x, y, u)$ . If  $|\{\delta_{(T; t)}(x, y, z), \delta_{(T; t)}(x, y, u), \delta_{(T; t)}(x, z, u), \delta_{(T; t)}(y, z, u)\}| = 1$ , then there is at least one vertex  $v$  on the path in  $T$  connecting  $\text{med}(x, y, z)$  and  $\text{med}(x, y, u)$  with  $t(v) \neq t(\text{med}(x, y, z))$ , as  $t$  is discriminating. Hence, there is a leaf  $e \in X$  such that  $v = \text{med}(x, y, e)$ . It follows that the set  $\{x, y, z, u, e\}$  is 4-6 partitioned by  $\delta_{(T; t)}$ , thus  $\delta_{(T; t)}$  satisfies Condition (\*).

On the other hand, if  $\delta$  is a symbolic ternary metric on  $X$  and satisfies (\*), then by Corollary 2, it corresponds to a unique complete quartet system, thus it encodes a binary phylogenetic tree  $T$  in view of Corollary 1. As in the last paragraph of the proof of Theorem 4, we can define a dating map  $t$  by  $t(v) = \delta(x, y, z)$  for every interior vertex  $v$  of  $T$  and for all 3-sets  $\{x, y, z\}$  whose median is  $v$ . Hence,  $(T, t)$  is a symbolically dated binary phylogenetic tree.  $\square$

## 5 The recognition of pseudo-cherries

In Theorem 4, we have established a 1-to-1 correspondence between symbolically dated phylogenetic trees and symbolic ternary metrics on  $X$ , and a bijection is given by mapping  $(T, t)$  to  $d_{(T, t)}$ . To get the inverse of this map, we can first compute the set of all quartets generated by a symbolic ternary metric, and then apply an algorithm that reconstructs a phylogenetic tree from the collection of all its displayed quartets. Finally, the dating map is defined as in our proof of Theorem 4. This approach would correspond to first extracting rooted triples from a symbolic ultrametric and then reconstruct the rooted tree (see Section 7.6 of [7]). However, a more direct way to reconstruct the corresponding tree from a symbolic ultrametric was presented in [5]. It is based on identifying maximal sets of at least two taxa that are adjacent to the same interior vertex, so-called *pseudo-cherries*. These can iteratively be identified into a single new taxon, thereby reconstructing the corresponding tree in a bottom-up fashion.

Here we show how to find the pseudo-cherries of  $T$  from a symbolic ternary metric  $d_{(T, t)}$ . Then an algorithm to reconstruct  $T$  can be designed exactly as in [5]. We point out that it is not necessary to check Condition (4) of a symbolic ternary metric, as a violation would make the algorithm recognize that the ternary map does not correspond to a tree.

Given an arbitrary symbolic ternary map  $\delta : X \times X \times X \rightarrow M^\odot$  satisfying Conditions (1), (2), and (3). For  $x, y \in X$  and  $m \in M^\odot$ , we say that  $x$  and  $y$  are *m-equivalent*, if there is  $z \in X$  such that  $\delta(x, y, z) = m$ , and for  $u, v \in X - x - y$ ,  $\delta(x, u, v) = m$  if and only if  $\delta(y, u, v) = m$ .

**Lemma 3.** *If  $x$  and  $y$  are  $m$ -equivalent and  $y$  and  $z$  are  $m'$ -equivalent, then  $m = m'$  and  $x$  and  $z$  are  $m$ -equivalent.*

*Proof.* Assume  $\delta(x, y, z) \neq m$ . Then let  $u \in X$  with  $\delta(x, y, u) = m$ . Since  $\delta$  is not constant on  $\{x, y, z, u\}$ , this 4-set must be 2-2-partitioned, thus exactly one of  $\delta(x, u, z) = m$  and  $\delta(y, u, z) = m$  must hold, in contradiction to  $x$  and  $y$  being  $m$ -equivalent. Hence, we have  $\delta(x, y, z) = m$ , and by symmetry we also have  $\delta(x, y, z) = m'$ , thus  $m = m'$ . In order to verify that  $x$  and  $z$  must be  $m$ -equivalent, we have already shown  $\delta(x, y, z) = m$ . For  $w, w' \in X - \{x, y, z\}$ , we have  $\delta(x, w, w') = m$  if and only if  $\delta(y, w, w') = m$ , since  $x$  and  $y$  are  $m$ -equivalent, and we have  $\delta(y, w, w') = m$  if and only if  $\delta(z, w, w') = m$ , since  $y$  and  $z$  are  $m$ -equivalent. Finally, we have  $\delta(x, y, w) = m$  if and only if  $\delta(x, z, w) = m$  if and only if  $\delta(y, z, w) = m$ . Hence  $x$  and  $z$  are  $m$ -equivalent.  $\square$

We say  $x, y \in X$  are  $\delta$ -equivalent, denoted by  $x \sim_\delta y$ , if there exists  $m \in M^\odot$  such that  $x$  and  $y$  are  $m$ -equivalent.

**Lemma 4.** *The relation of being  $\delta$ -equivalent is an equivalence relation.*

*Proof.* For any  $x \in X$ , since  $\delta(x, x, y) = \odot$  for any  $y \in X$ , by definition  $x$  and  $x$  are  $\odot$ -equivalent, hence  $x$  and  $x$  are  $\delta$ -equivalent. Hence  $\sim_\delta$  is reflexive. For any  $x \sim_\delta y$ , we know that there exists an  $m \in M^\odot$  such that  $x$  and  $y$  are  $m$ -equivalent. Since  $\delta$  is symmetric, by the definition of  $m$ -equivalent,  $y$  and  $x$  are also  $m$ -equivalent, thus  $y \sim_\delta x$ . Hence,  $\sim_\delta$  is symmetric. To prove the transitivity of  $\sim_\delta$ , assume  $x \sim_\delta y$  and  $y \sim_\delta z$ , by Lemma 3 we know that  $x \sim_\delta z$ . Therefore,  $\delta$ -equivalent is an equivalence relation.  $\square$

Suppose that  $T$  is a phylogenetic tree on  $X$ . Let  $C \subseteq X$  be a subset of  $X$  with  $|C| \geq 2$ . We call  $C$  a *pseudo-cherry* of  $T$ , if there is an interior vertex  $v$  of  $T$  such that  $C$  is the set of all leaves adjacent to  $v$ .

**Theorem 7.** *If  $(T; t)$  is a symbolically dated phylogenetic tree, then a non-empty subset  $C$  of  $X$  is a non-trivial equivalence class of  $\sim_{\delta_{(T, t)}}$  if and only if  $C$  is a pseudo-cherry of  $T$ .*

*Proof.* For the ease of notation, we let  $\delta = \delta_{(T, t)}$ .

Since  $t$  is discriminating, the definition of a pseudo-cherry immediately implies that any pseudo-cherry of  $T$  must be contained in a non-trivial equivalence class of  $\sim_\delta$ .

Conversely, if a non-trivial equivalence class  $C$  of  $\sim_\delta$  is not a pseudo-cherry, then there are  $x_1, x_2 \in C$  such that the path in  $T$  that contains  $x_1$  and  $x_2$  has length at least 3, and since  $t$  is discriminating, it has at least 2 interior vertices labeled by two different elements of  $M$ . Suppose that all elements of  $C$  are  $m$ -equivalent, and that  $v$  is an interior vertex on the path from  $x_1$  to  $x_2$  such that  $t(v) = m'$  and  $m' \neq m$ . Further, let  $y \in X$  such that  $v = \text{med}(x_1, x_2, y)$ . Since  $x_1$  and  $x_2$  are  $m$ -equivalent, there is

$z \in X$  such that  $\delta(x_1, x_2, z) = m$ . Then the median  $u$  of  $x_1, x_2, z$  is also on the path from  $x_1$  and  $x_2$ , and we assume without loss of generality that  $u$  is on the path from  $x_1$  to  $v$ . It follows that  $\delta(x_1, y, z) = m$  but  $\delta(x_2, y, z) = m'$  in contradiction to  $x_1 \sim_m x_2$ .  $\square$

## 6 Discussions and open questions

The proofs of our main results heavily rely on extracting the corresponding quartet set from a symbolic ternary metric and then checking that our Conditions (3) and (4) guarantee the quartet system to be thin, transitive, and saturated, and adding (\*) makes the quartet system complete. The conditions look like (3) corresponds to thin and transitive, (4) to saturated, and (\*) to complete. However, this is not true, and removing (4) from Theorem 5 does not necessarily yield a transitive complete quartet system. While for five taxa, a 5-5-partition yields the only non-saturated complete transitive quartet system, Lemma 2 does not hold without Condition (4). Indeed the ternary map that is visualised in Figure 5 satisfies Conditions (1), (2), (3), and (\*), but it generates two quartets on each of  $\{a_1, a_2, b_1, b_2\}$  and  $\{a_1, a_2, c_1, c_2\}$ . It can be shown by checking the remaining 5-sets in Case 2 of our proof of Lemma 2 that every ternary map on 6 taxa satisfying Conditions (1), (2), (3), (\*) that does not yield a thin quartet system is isomorphic to this example. This raises the question whether ternary maps satisfying these four conditions can be completely characterised. The hope is to observe something similar to the Clebsch trees that were observed by Jan Weyer-Menkhoff [8]. As a result, a phylogenetic tree with all interior vertices of degree 3, 5, or 6 can be reconstructed from every transitive complete quartet set.

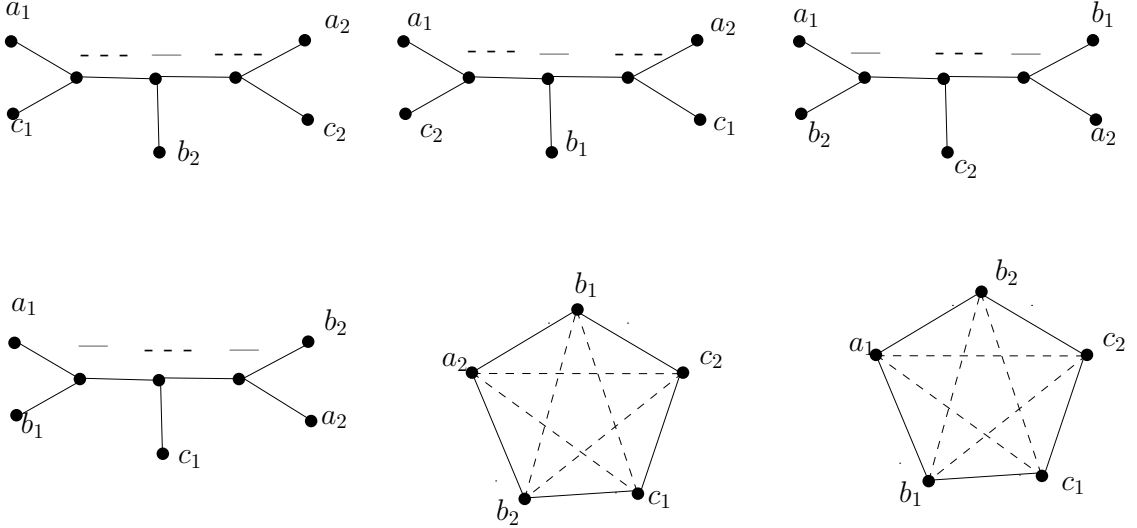


Figure 5: The 5-taxon trees respectively graph representations generated by a ternary map satisfying Conditions (1), (2), (3), (\*) but not (4).

Another direction to follow up this work would be to consider more general graphs than trees. A *median graph* is a graph for which every three vertices have a unique median. Given a vertex-coloured median graph and a subset  $X$  of its vertex set, we can get a symmetric ternary map on  $X \times X \times X$  by associating the colour of the median to every 3-subset of  $X$ . It would be interesting to see whether this map can be used to reconstruct the underlying graph for other classes of median graphs than phylogenetic trees. In phylogenetics, median graphs are used to represent non-treelike data. Since the interior vertices of those so-called *splits graphs* do in general not correspond to any ancestor of some of the taxa, reconstructing a collection of splits from the ternary map induced by a vertex-coloured splits graph is probably limited to split systems that are almost compatible with a tree.

It is one of the main observations of [5] that the 4-point condition for symbolic ultrametrics can be formulated in terms of *cographs* which are graphs that do not contain an induced path of length 3. For the special case that  $\delta : X \times X \rightarrow M^\odot$  with  $|M| = 2$  and  $m \in M$ , consider the graph with vertex set  $X$  where two vertices  $x, y$  are adjacent, if and only if  $\delta(x, y) = m$ . Then deciding whether  $\delta$  is a symbolic ultrametric can be reduced to checking whether this graph (as well as its complement) is a cograph. This is useful for analysing real data which will usually not provide a perfect symbolic ultrametric, thus some approximation is required. For ternary maps and unrooted trees, the 5-taxon case looks promising, as Condition (4) translates to a forbidden graph representation that splits the edges of a  $K_5$  into two 5-cycles, thus we have a self-complementary forbidden induced subgraph. However, for more taxa, the 3-sets that are mapped to the same value of a ternary map  $\delta$  define a 3-uniform hypergraph on  $X$  and formulating Condition (4) in terms of this hypergraph does not seem to be promising. In addition, even if there are only two values of  $\delta$  for 3-sets, Condition (3) does not become obsolete. We leave it as an open question, whether an alternative characterisation of symbolic ternary metrics exists that makes it easier to solve the corresponding approximation problem.



A possible application of symbolic ternary metrics is to analyse orthology relations. The taxa are homologous genes, and the true underlying structure is a gene tree where every interior vertex either corresponds to a speciation or a duplication. Two genes are called orthologs if their last common ancestor corresponds to a speciation (and paralogs otherwise). The orthology relation can be estimated by sequence comparison. Following [5], this can be used to construct a gene tree for every three taxa, including assigning either a speciation or a duplication to their median. This defines a ternary map on the set of all genes, and if there is no error in the orthology relation, then it is a symbolic ternary metric. It corresponds to the tree observed from the true gene tree by identifying all edges for which both incident vertices correspond to the same kind of event. Therefore, an unrooted analogue to the tree reconstruction in [6] is possible.

## Acknowledgements

We thank Peter F. Stadler for suggesting to consider general median graphs. This work is supported by National Natural Science Foundation of China (No. 11271255 and No. 11671258). YL acknowledges support of Postdoctoral Science Foundation of China (No. 154463).

## References

- [1] Hans-Jürgen Bandelt and Michael Anthony Steel. Symmetric matrices representable by weighted trees over a cancellative abelian monoid. *SIAM Journal on Discrete Mathematics*, 8(4):517–525, 1995.
- [2] Sebastian Böcker and Andreas WM Dress. Recovering symbolically dated, rooted trees from symbolic ultrametrics. *Advances in mathematics*, 138(1):105–125, 1998.
- [3] Hans Colonius and Hans Hennig Schulze. Tree structures for proximity data. *British Journal of Mathematical and Statistical Psychology*, 34(2):167–180, 1981.
- [4] Andreas Dress, Katharina T Huber, and Jacobus Koolen. *Basic phylogenetic combinatorics*. Cambridge University Press, 2012.
- [5] Marc Hellmuth, Maribel Hernandez-Rosales, Katharina T Huber, Vincent Moulton, Peter F Stadler, and Nicolas Wieseke. Orthology relations, symbolic ultrametrics, and cographs. *Journal of mathematical biology*, 66(1-2):399–420, 2013.
- [6] Marc Hellmuth, Nicolas Wieseke, Marcus Lechner, Hans-Peter Lenhof, Martin Middendorf, and Peter F. Stadler. Phylogenetics from paralogs. *Proc. Natl. Acad. Sci. USA*, 112:2058–2063, 2015. doi: 10.1073/pnas.1412770112.
- [7] Charles Semple and Mike A Steel. *Phylogenetics*, volume 24. Oxford University Press on Demand, 2003.
- [8] Jan Weyer-Menkhoff. *New quartet methods in phylogenetic combinatorics*. PhD thesis, Universität Bielefeld, 2003.